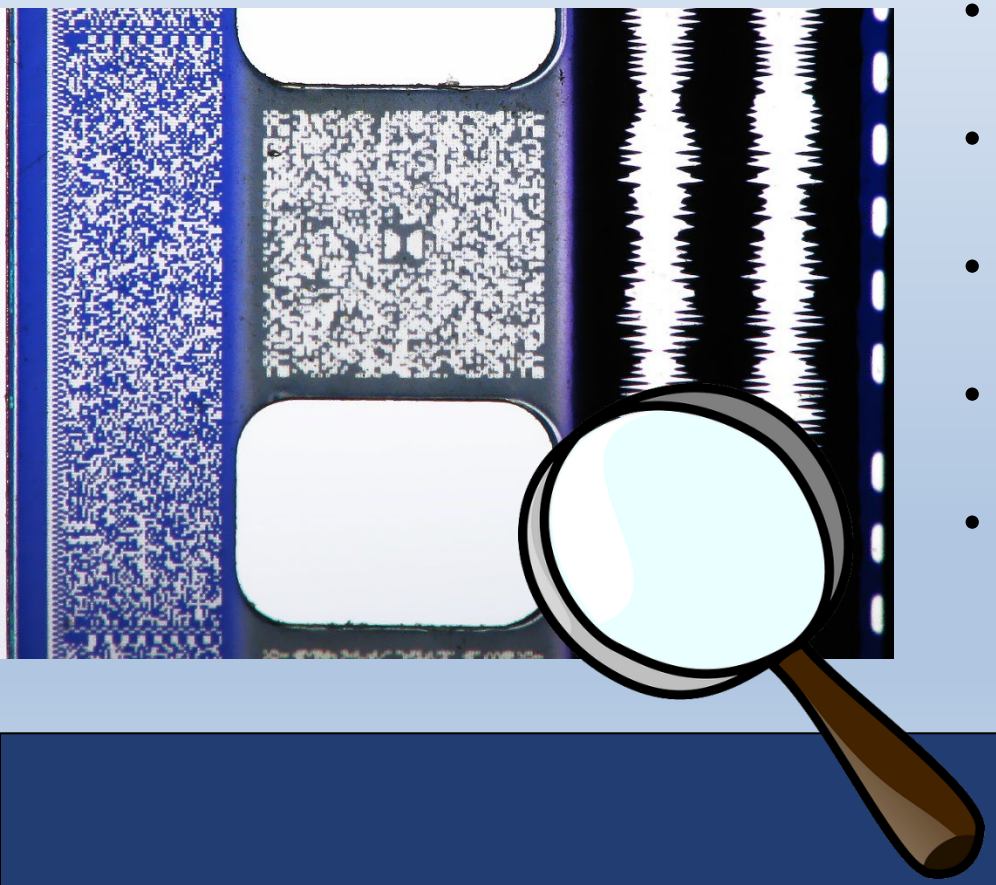
Tecnologías para el análisis y extracción de metada en contenidos audiovisuales: Tecnologías del Habla

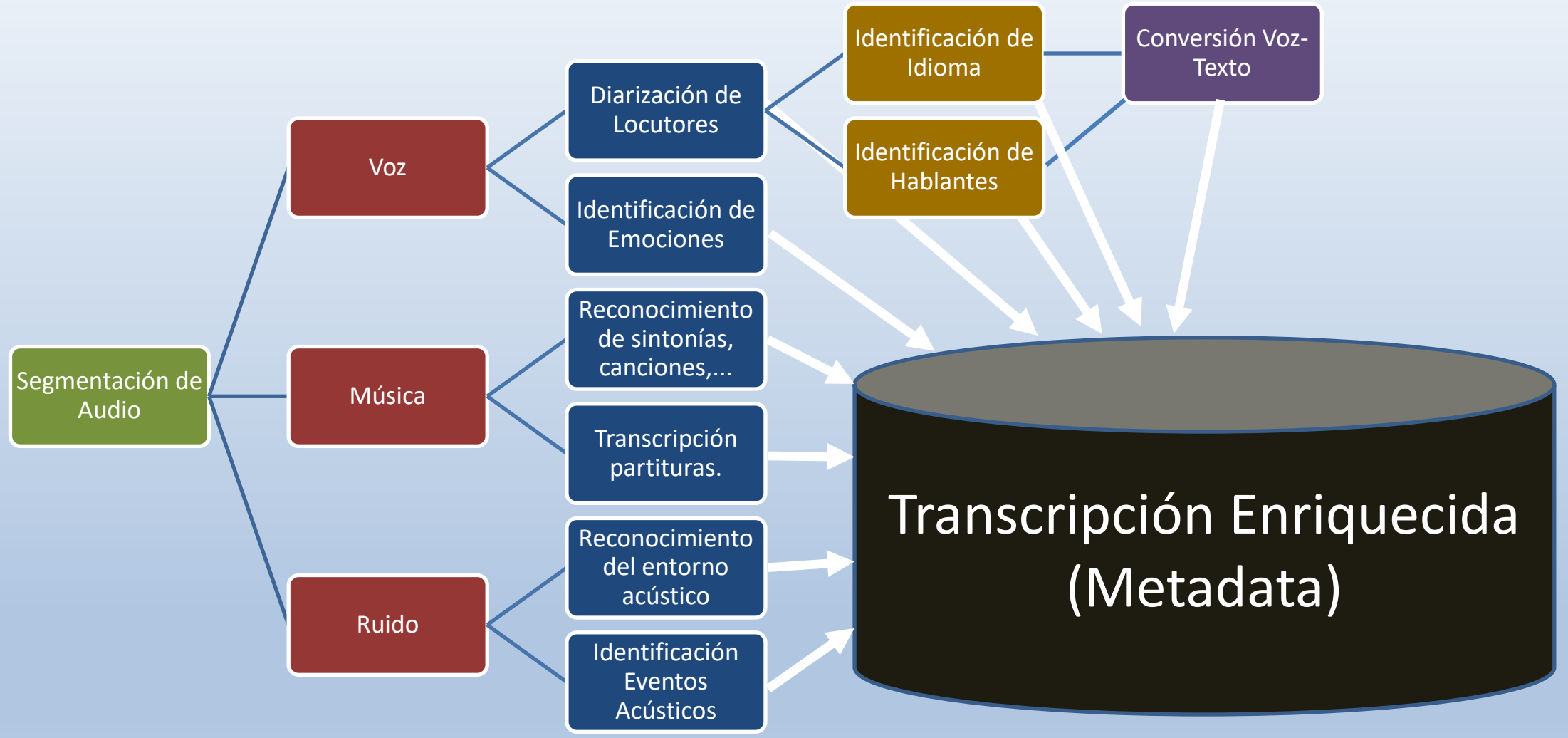
Extracción de Información: Audio

¿Qué información podemos encontrar en un audio?

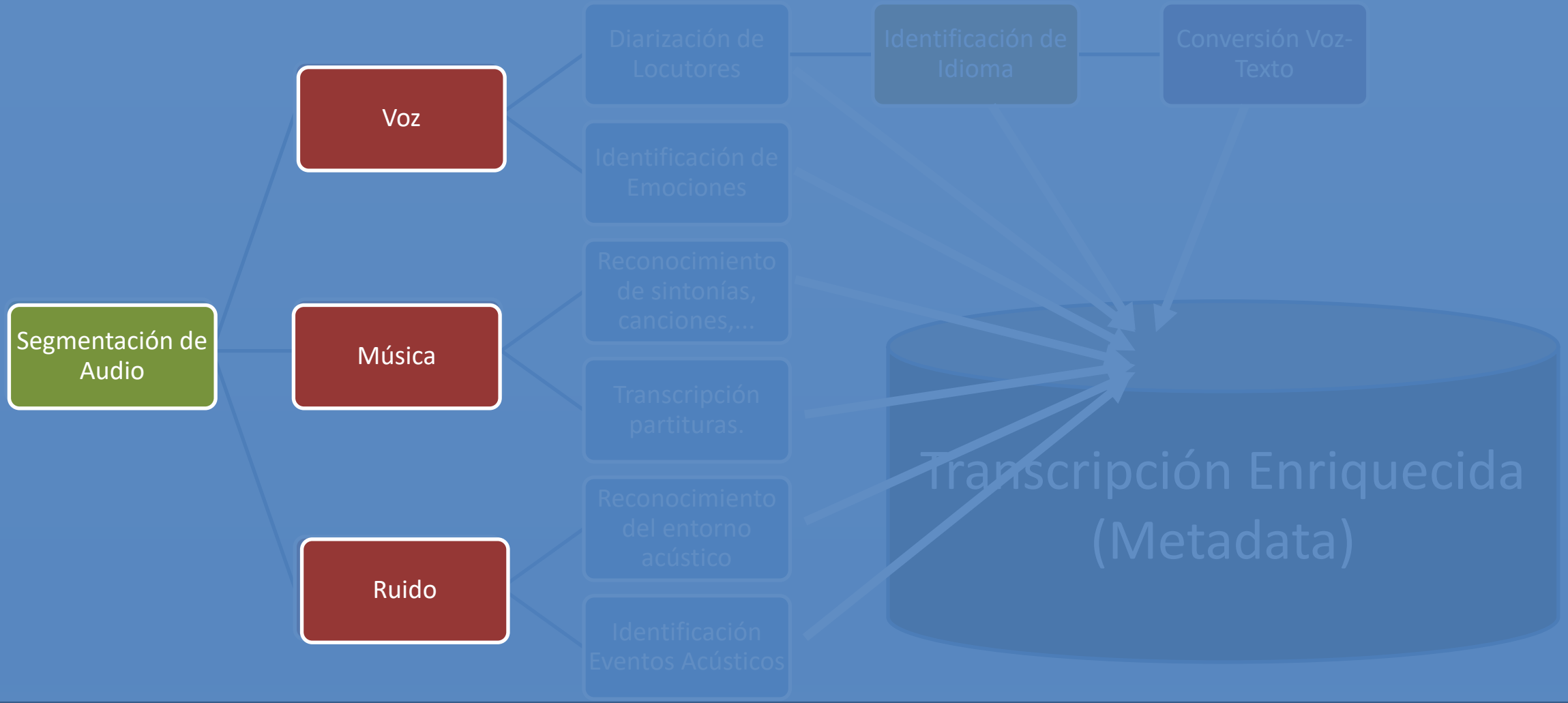


- Hay ruido, música, habla, …
- Cuántas personas hablan y cuándo habla cada una de ellas
- Cuáles son las identidades de las personas que hablan
- En qué idioma están hablando
- Qué dice cada una de ellas
- Cuál es el estado emocional de cada una de ellas

Tecnologías



Segmentación de Audio



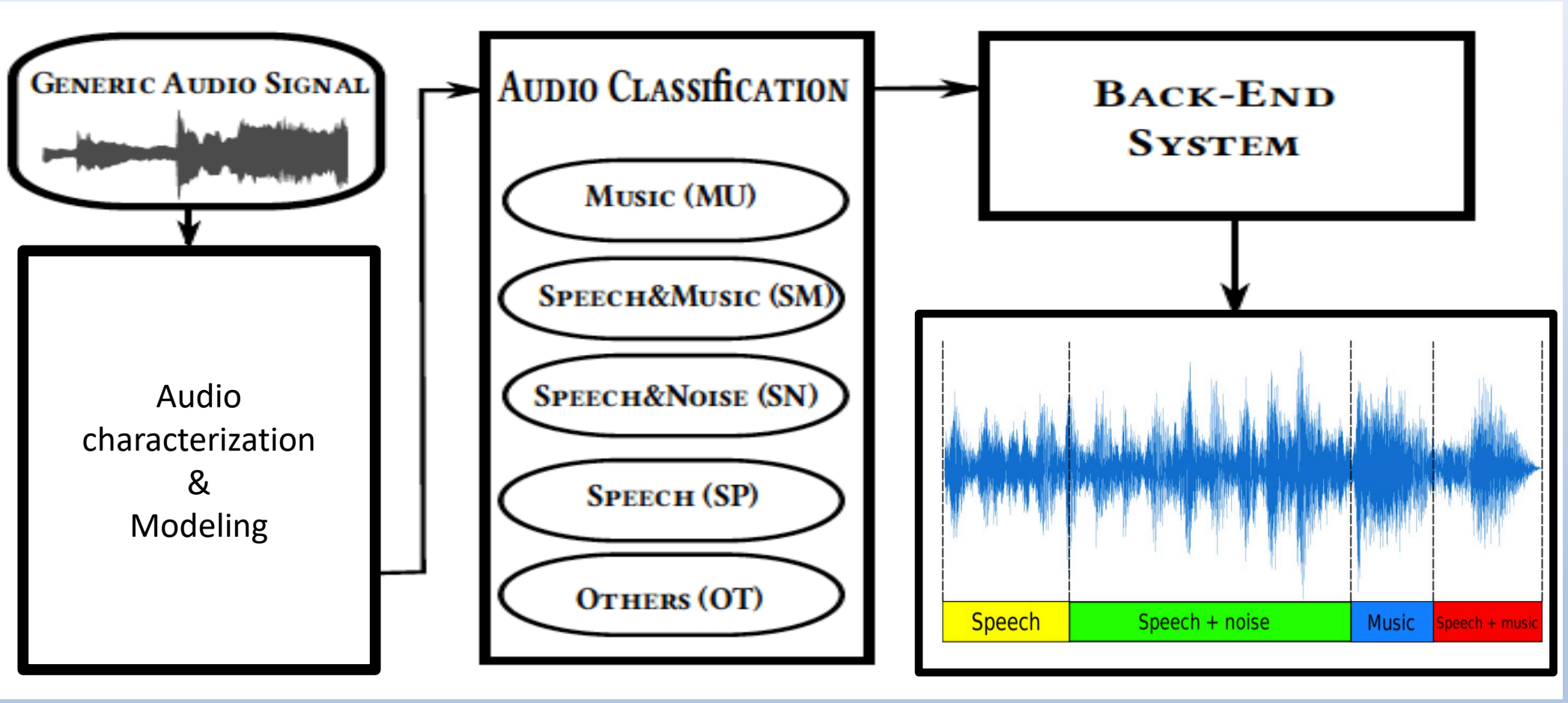
Segmentación de Audio

- **¿Qué es?:**
  - **Dividir el audio de entrada en fragmentos atendiendo al tipo de contenido acústico: Voz / Música / Ruido y combinaciones de estos.**

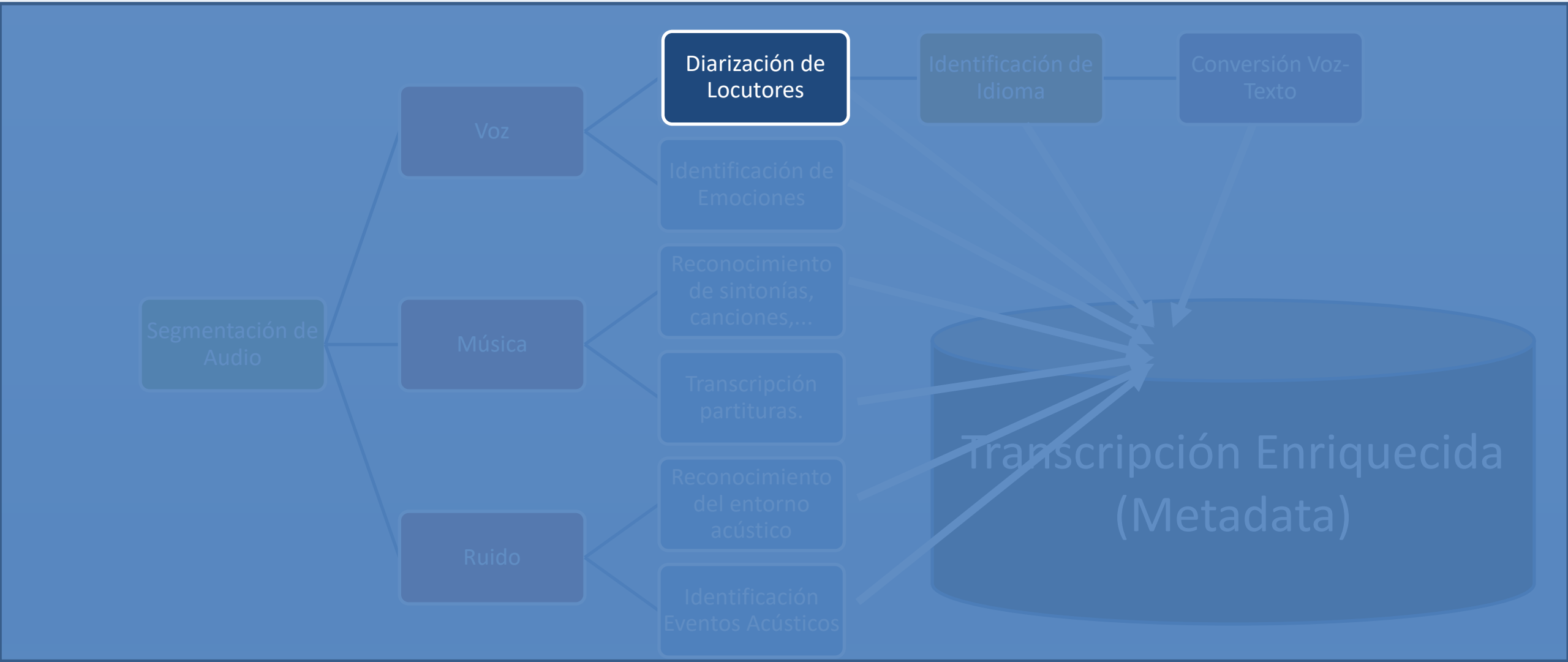
Segmentación de Audio

- **¿Para qué sirve?:**
  - Da soporte a otras tareas de extracción de información como:
    - Diarización
    - Identificación del hablante
    - Conversión Voz-Texto
    - …

Segmentación de Audio



Tecnologías



Segmentación y Agrupación de Hablantes

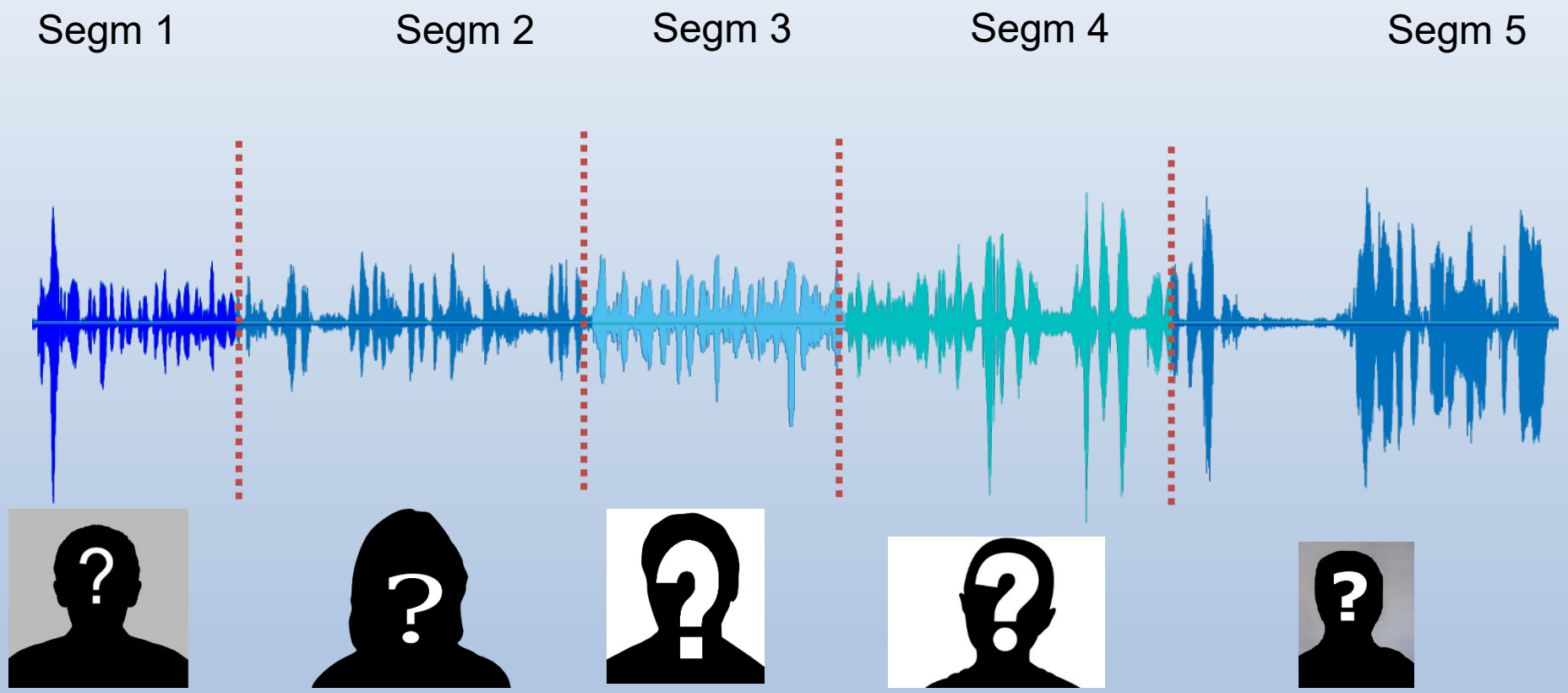
- **<u>¿Qué es?:</u>**

  - **Dividir en fragmentos atendiendo al interviniente y agrupar dichos fragmentos en función de la identidad del locutor.**

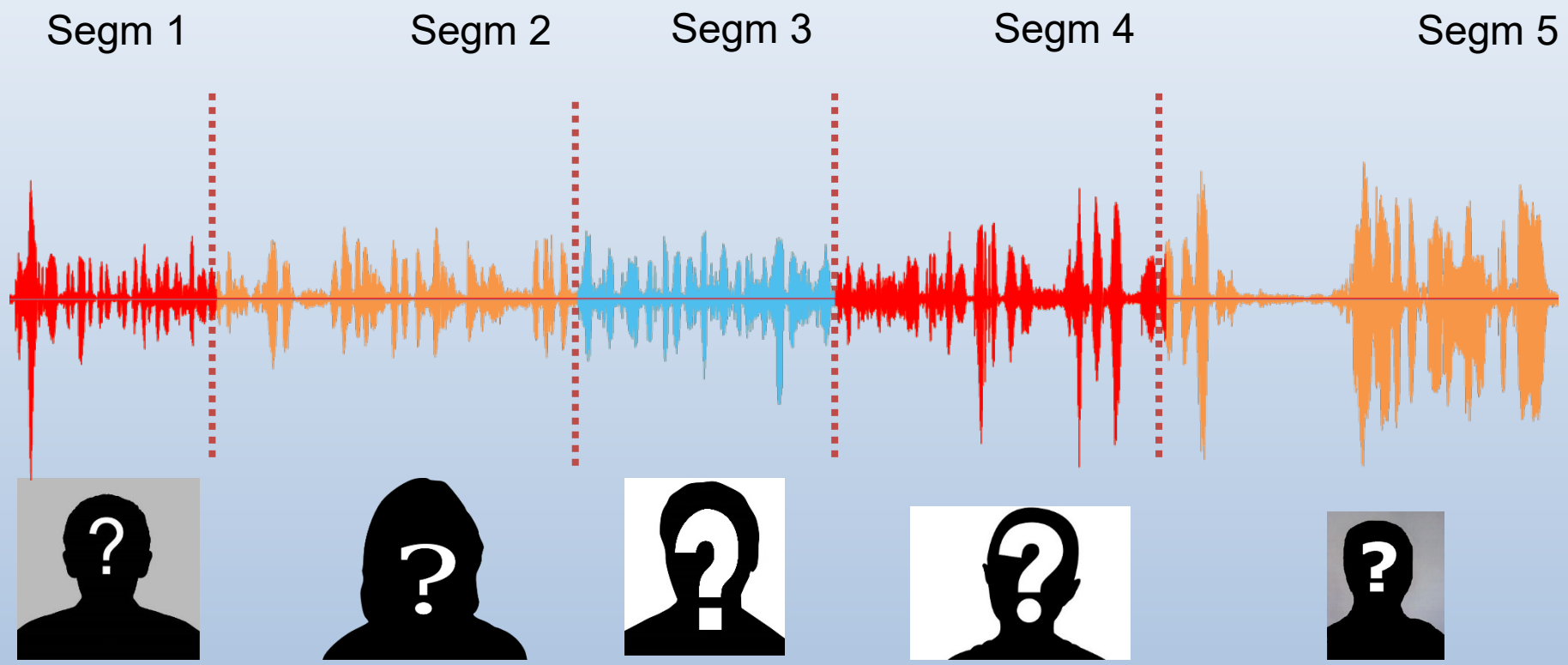
  - **Término usado por la comunidad: Diarización**
    - *Diarise: (Diarize) to make use of a diary to record past events or those planned for the future.*

Segmentación y Agrupación de Hablantes

Segmentación y Agrupación de Hablantes

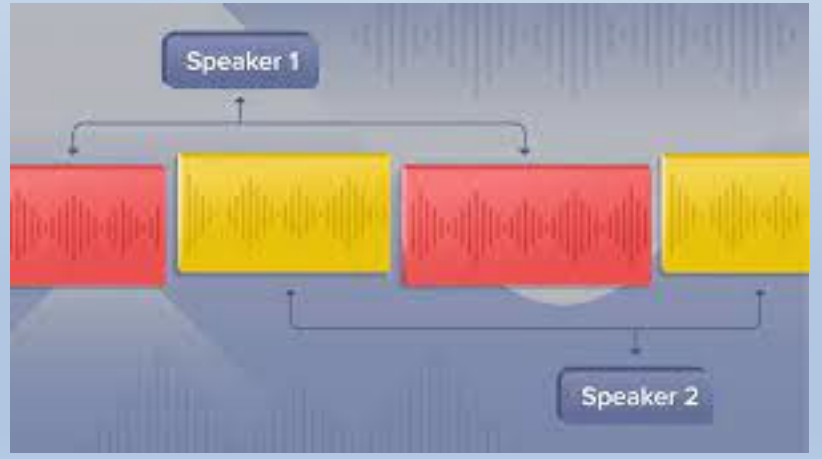


Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

- **¿Para qué sirve?:**
  - Tecnología soporte para mejorar prestaciones de:
    - Reconocimiento automático del habla
    - Reconocimiento del hablante
    - …

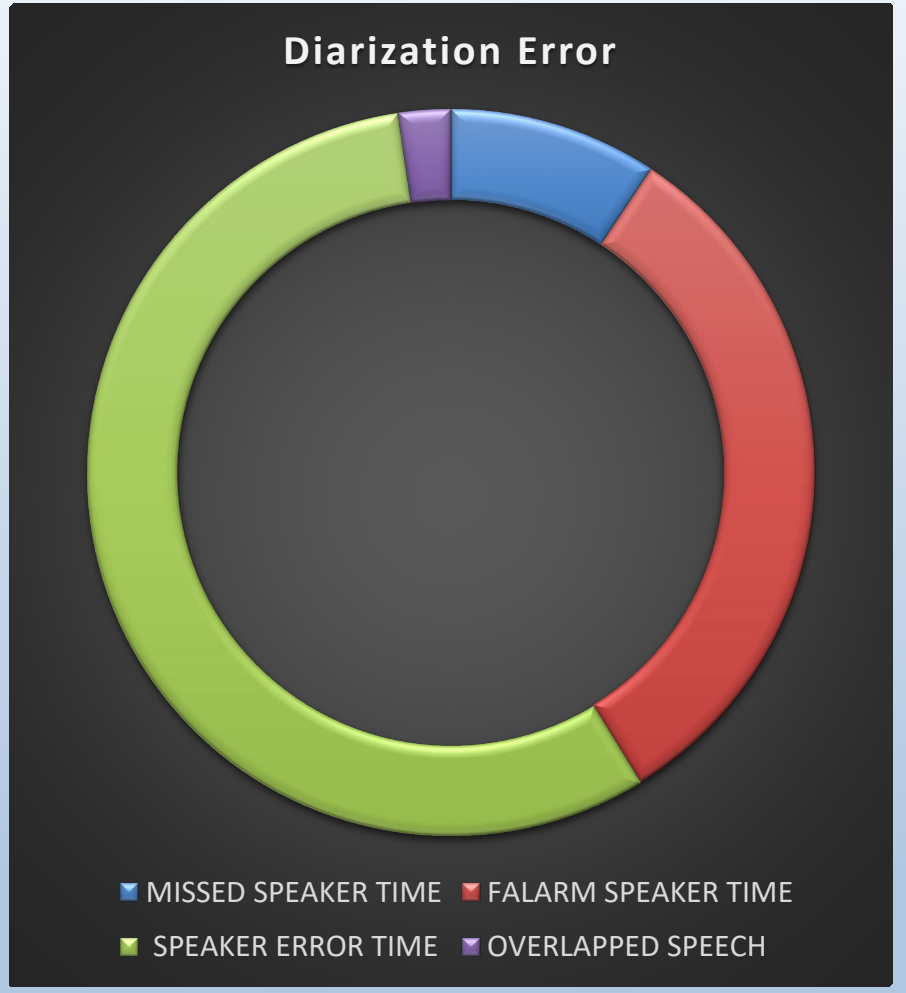


Medida del Error

- **Diarization Error Rate:**

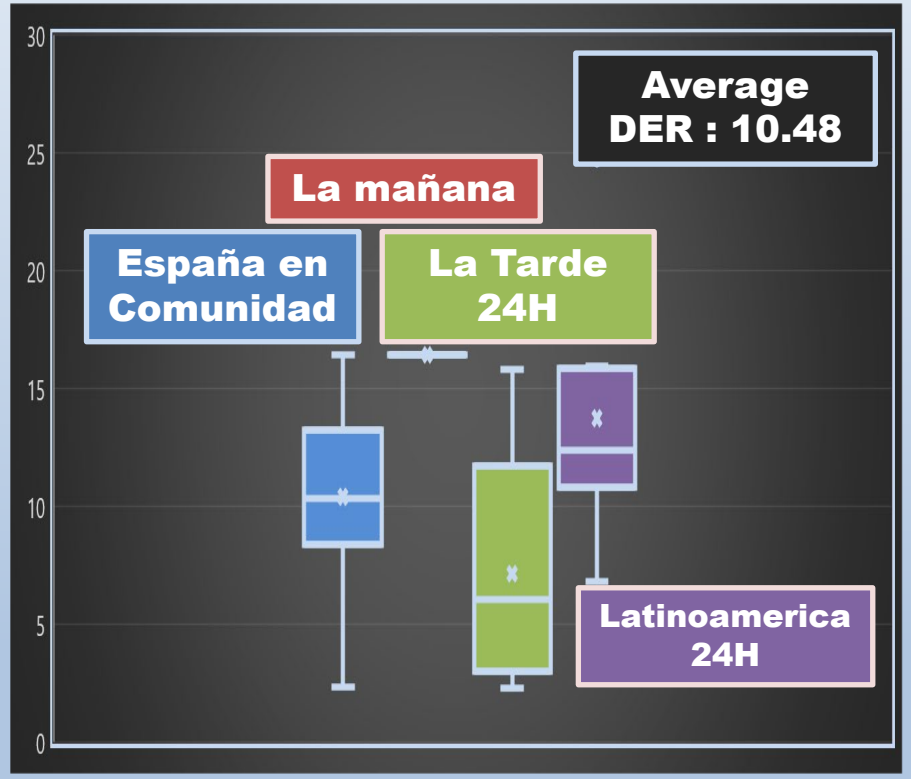
$$DER = \frac{T_{incorrecto}}{T_{voz}}$$

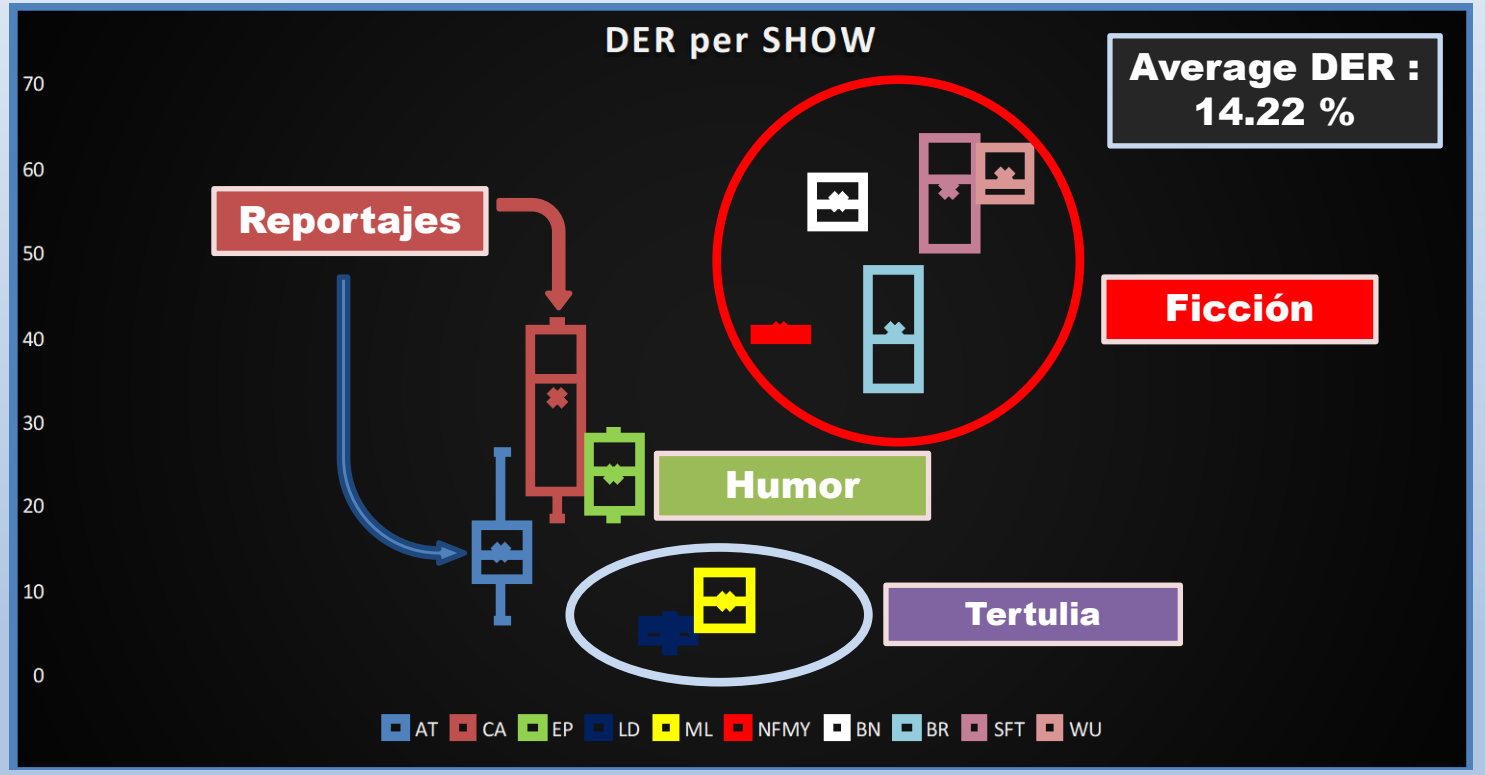
  - Componentes del error:
    - Pérdida:
      - Hay voz pero se ha confundido con silencio.
    - Falsa Alarma:
      - No hay voz pero se ha detectado erróneamente.
    - Error de Locutor:
      - Se ha confundido un locutor con otro.
    - Error por Solape:
      - Dos locutores hablan a la vez, pero solo se ha identificado a uno.



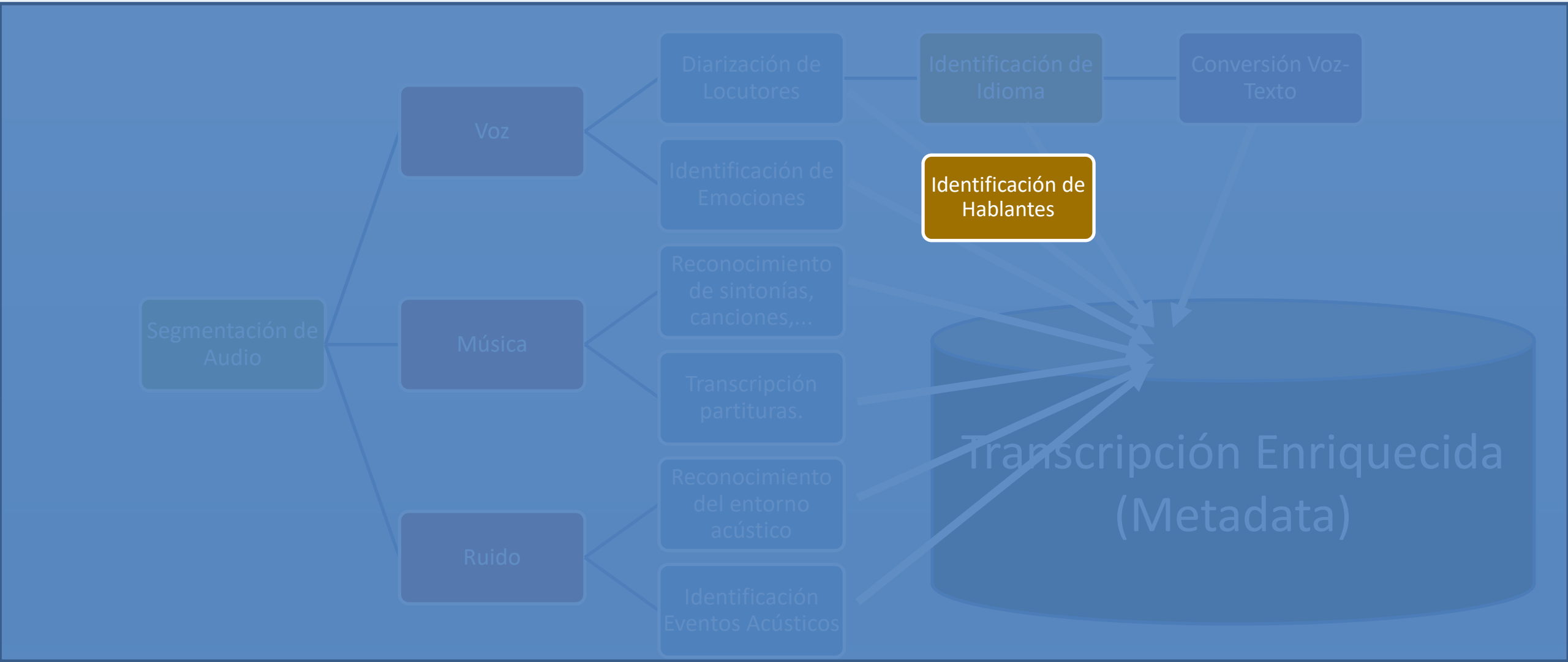
Prestaciones: Albayzin 2018 y 2020

- **Broadcast:**





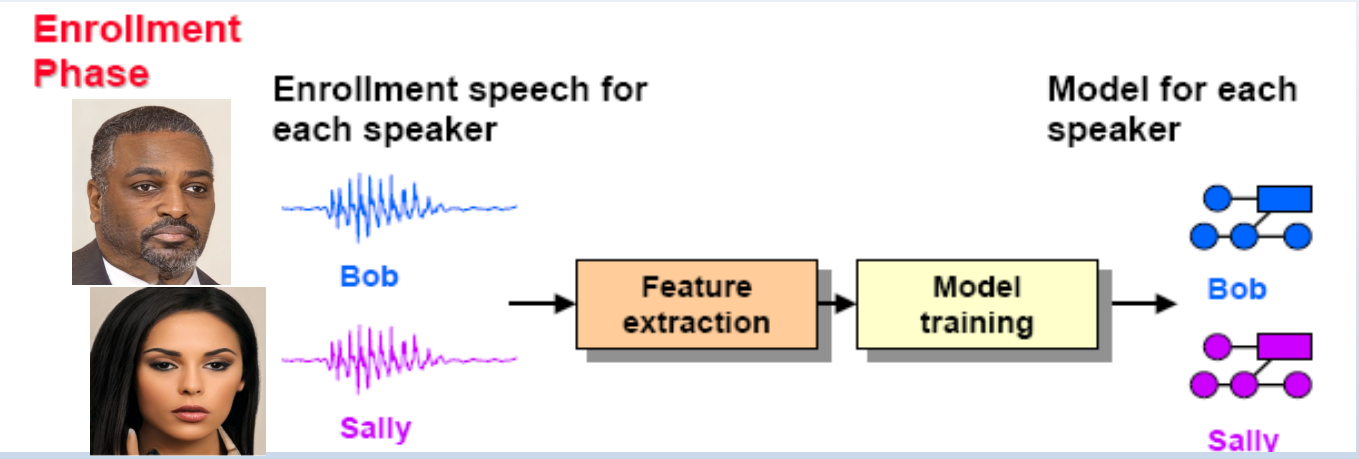
Tecnologías

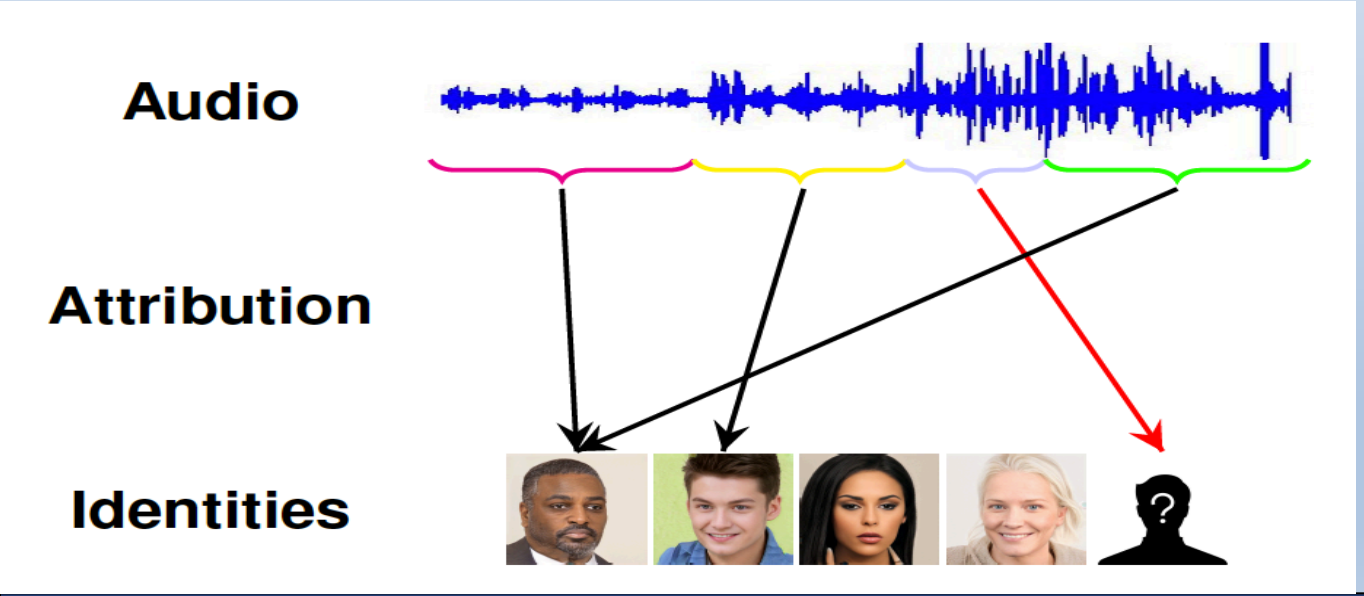


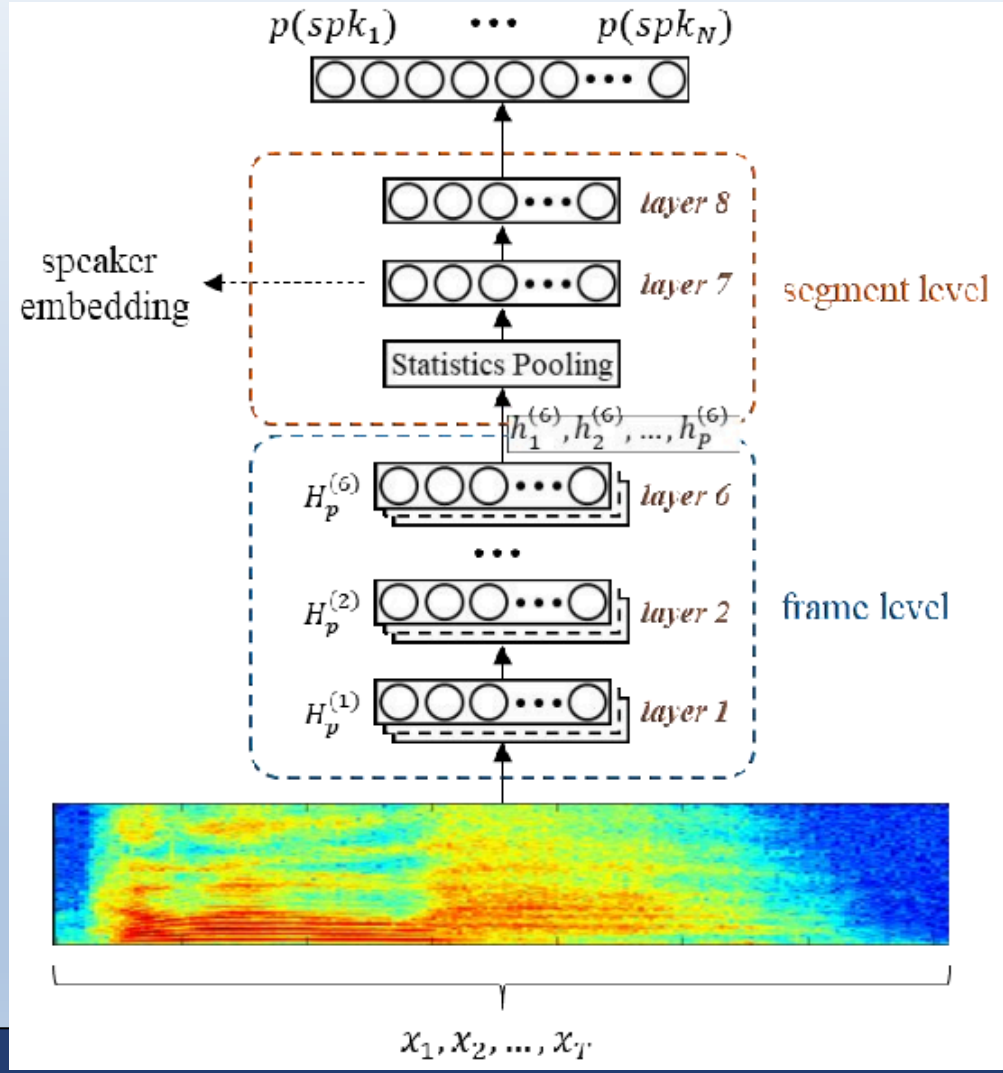
Identificación de Hablantes

- **¿Para qué sirve?:**
  - **Permite asignar identidades concretas a fragmentos de audio de un contenido analizado**

Identificación de Hablantes







Viñals, Ignacio; Ortega, Alfonso; Miguel, Antonio; Lleida, Eduardo
The Domain Mismatch Problem in the Broadcast Speaker Attribution Task
Applied Sciences, vol. 11, no. 18, pp. 8521, 2021.

VIVOLAB

Universidad Zaragoza

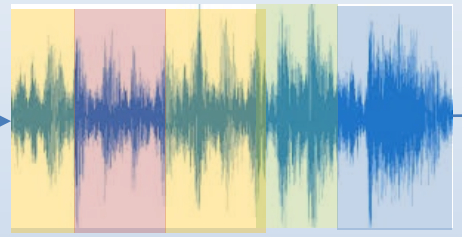
Identificación de Hablantes

- Reconocimiento de personajes en programas de TV:

Extract Audio and Content

Diarize Audio

Spk A Spk A Spk D



Spk B Spk C

Assign Speaker Identities

Speaker Verification Engine

Enroll Speakers

SPK Database

Spk Metadata File

FileA John Smith start1 end1
File A John Smith start2 end2
File B John Smith start1 end1
File C John Smith start 1 end1
FileB Mary Adams start1 end 1
File C Sam Sanders start1 end 1

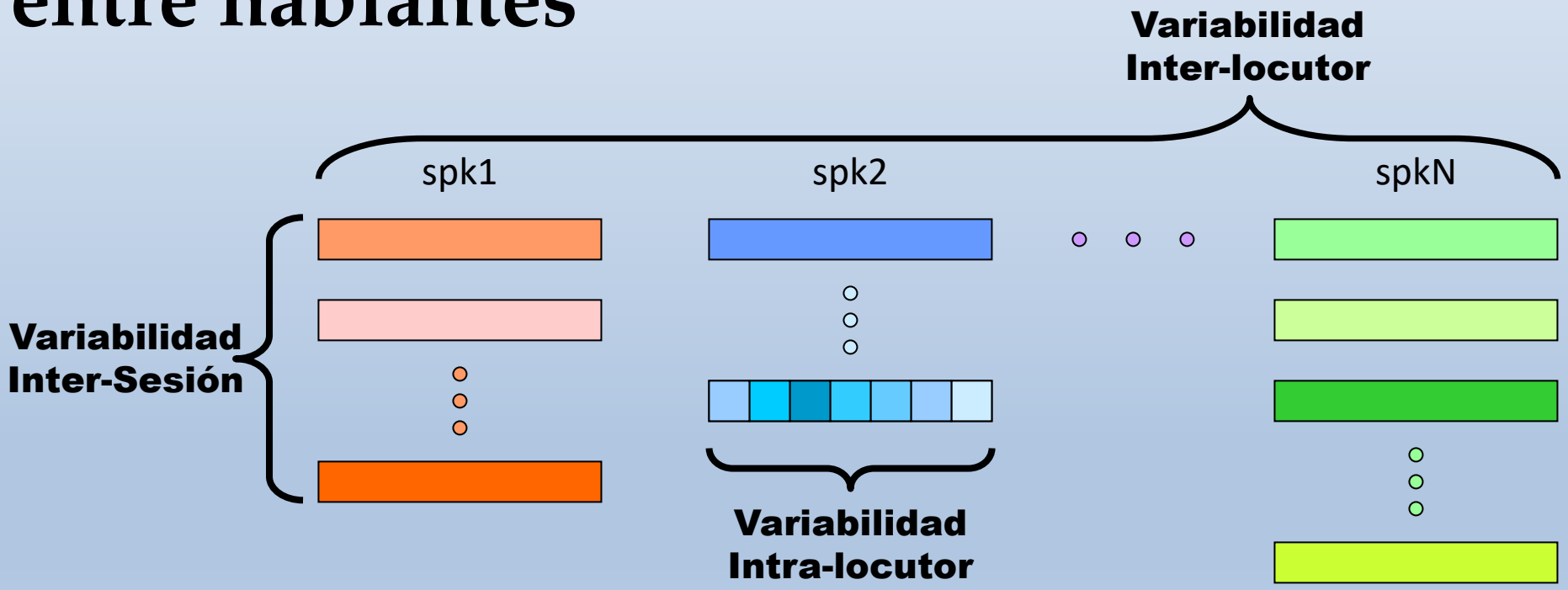
...

VIVOLAB

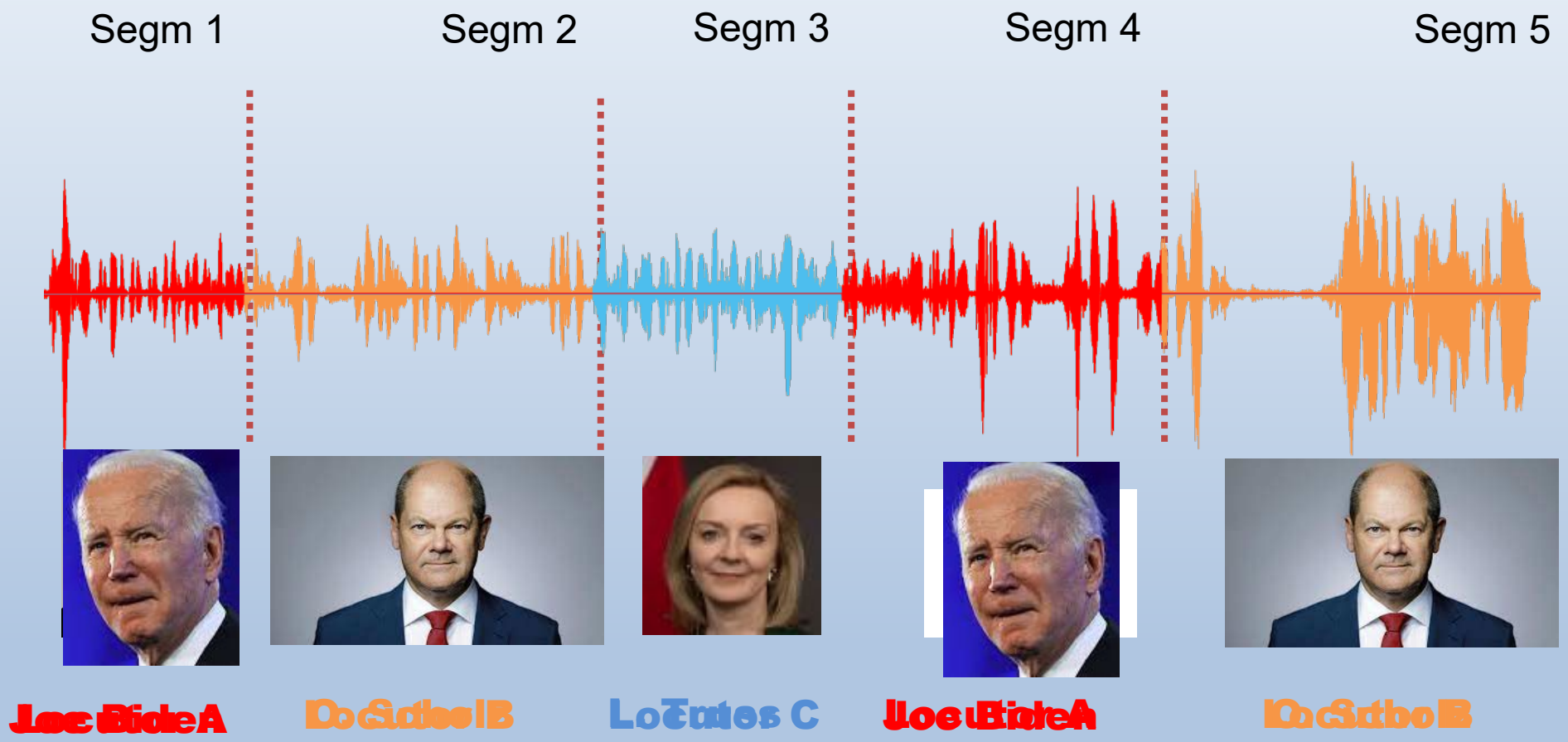
Universidad Zaragoza

Retos en la Identificación de hablantes

- **Alta variabilidad en la voz de los hablantes**
- **Diversidad de dominios acústicos**
- **Solape entre hablantes**



Diarización Junto con Identificación de Hablante:



Prestaciones: Albayzin 2020

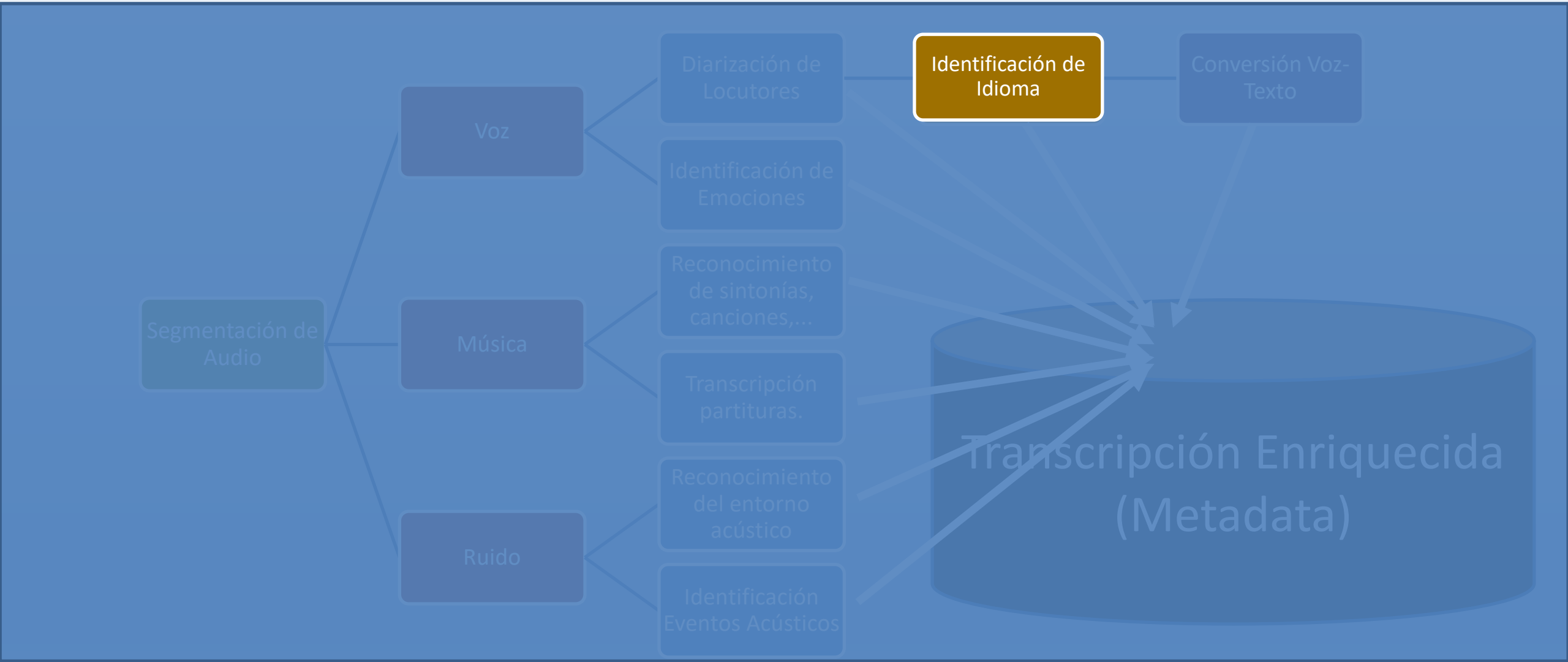
- *Iberspeech-RTVE-Challenge* :

	AER
BIOMETRIC VOX	65.09 %
VIVOLAB	72.63 %

	MISS	FA	SPK ERR
BIOMETRIC VOX	47.0 %	9.2 %	8.9 %
VIVOLAB	5.1 %	53.3 %	14.2 %

Subset	Closed Condition			Open Condition		
	Direct	Indirect	Hybrid	Direct	Indirect	Hybrid
Dev. subset	13.73	15.27	15.89	41.91	**37.45**	37.68
Eval. subset	25.11	17.20	**16.49**	65.31	60.34	**31.95**

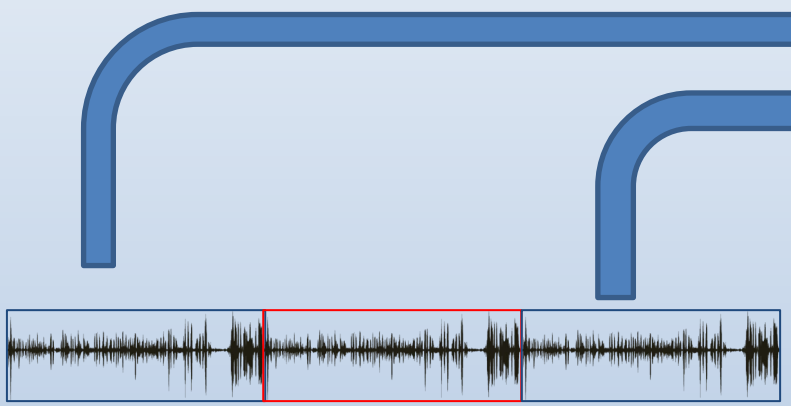
Tecnologías



Identificación de Idioma

- **¿Para qué sirve?:**
  - **En entornos multilingüe, permite el indexado y la recuperación de documentos:**
  - **Esencial en esos entornos como soporte a:**
    - **Reconocimiento automático del habla**

Identificación de Idioma



Conversión Voz-Texto (Castellano)

Conversión Voz-Texto (Catalán)

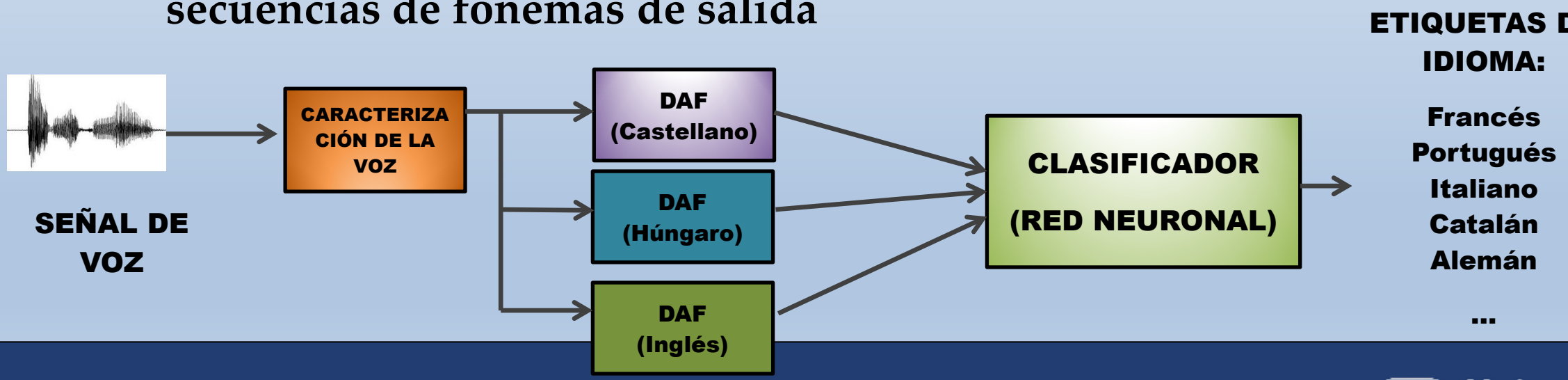
La Fundación ”la Caixa” ha hecho balance de la asistencia a sus centros y exposiciones culturales y científicas durante 2021, año todavía marcado por la pandemia. La entidad ha hecho un gran esfuerzo para recuperar el ritmo de visitantes en los ocho centros CaixaForum y en CosmoCaixa, y en las exposiciones itinerantes, para conseguir que la cultura sea segura siguiendo las restricciones de capacidad para el cóvido estipuladas por las comunidades autónomas. El público ha respondido y 5,2 millones de personas han visitado los centros y han vivido las diferentes propuestas culturales, científicas y educativas. Son 1,3 millones de personas más que en 2020 (+35,7%).

Així, set de cada deu visitants habituals abans de la pandèmia han tornat als centres CaixaForum i CosmoCaixa, un total d’1,9 milions de persones. Cal sumar-hi els 3,3 milions de visitants de les exposicions itinerants en més de 70 ciutats de tot Espanya.

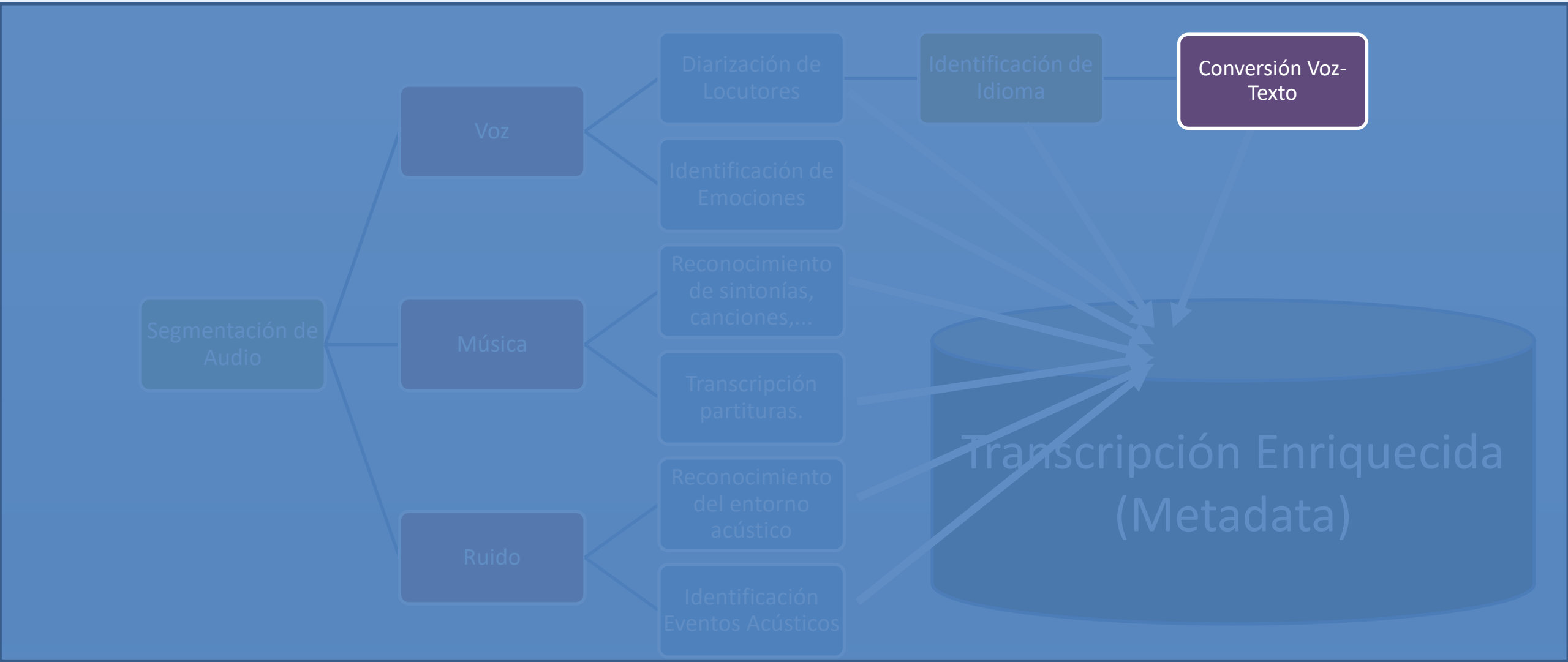
CaixaForum Barcelona, la més visitada, ha sido la exposición interactiva e inmersiva Teamlab. Arte, tecnología y naturaleza (en la imagen), seguida de El sueño americano. Del pop a la actualidad, con obras de Andy Warhol o Roy Lichtenstein, producida con la colaboración del British Museum. Por su parte, CosmoCaixa ha acogido, con gran afluencia de público, PRINT3D. Reimprimir la realidad y Nikola Tesla. El genio de la electricidad moderna, esta última vigente hasta febrero de 2022.

Identificación de Idioma

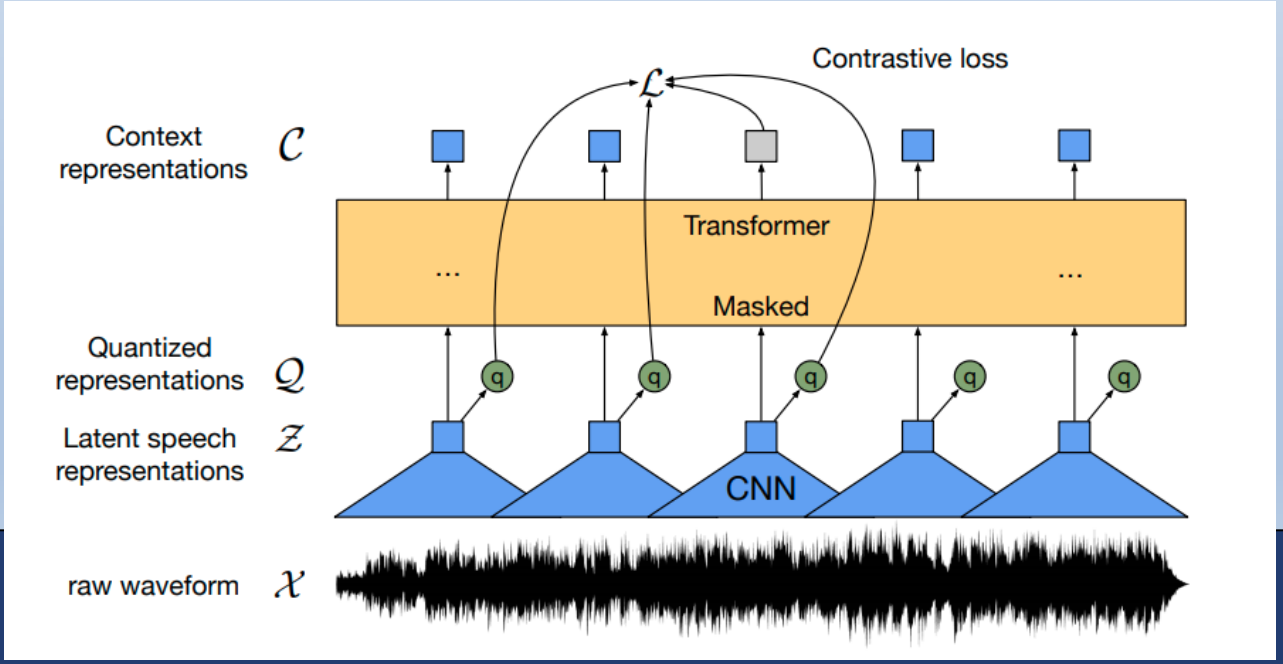
- **Tecnologías:**
  - **Acústicos**
    - **Tratan de buscar patrones discriminativos directamente sobre la señal de voz**
  - **Fonotácticos (Lingüísticos)**
    - **Primero procesan la señal de entrada con un (o varios) reconocedor fonético (en varios idiomas) después buscan patrones discriminativos en las secuencias de fonemas de salida**

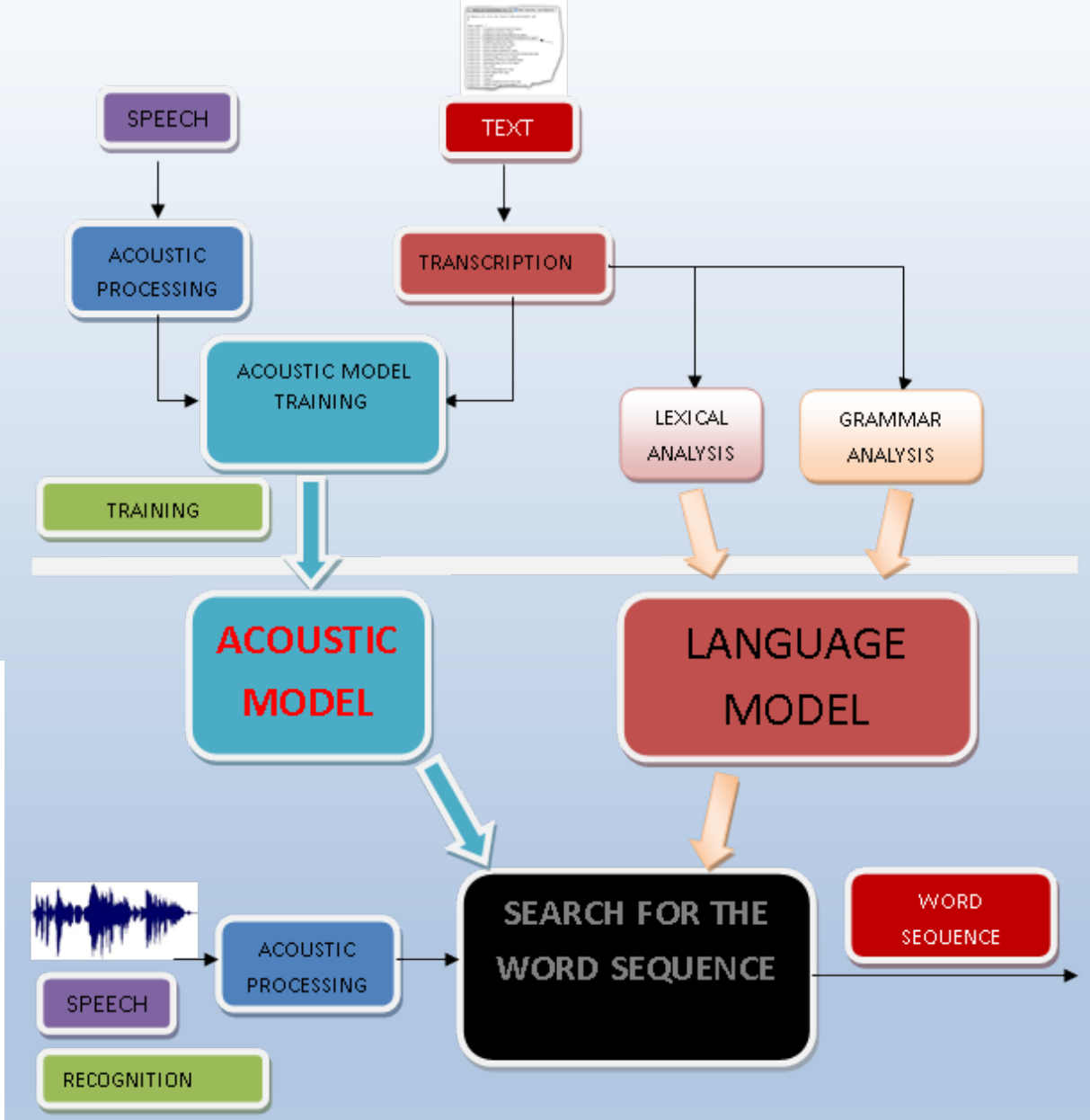


Tecnologías



Etapas y Procesos Reconocimiento Automático del Habla:





Componentes de un Sistema de RAH

- **MODELO ACÚSTICO**
  - Describe las características de cada unidad desde el punto de vista de la señal de voz (espectralmente)
- **MODELO DE LENGUAJE**
  - Describe las relaciones entre palabras del vocabulario
  - Cuantifica la probabilidad de las secuencias de palabras
- **MODELO LÉXICO**
  - Describe cómo se forma cada palabra del vocabulario a partir de las diferentes unidades del modelo acústico.

ERRORES EN UN SISTEMA RAH

- **Borrados**
  - El locutor dice algo pero el sistema no devuelve nada
- **Substituciones**
  - El sistema devuelve a su salida una palabra diferente de la pronunciada por el locutor.
- **Inserciones**
  - El locutor no dice nada, pero el sistema devuelve alguna palabra (generalmente debido a artefactos acústicos)

ERRORES EN UN SISTEMA RAH

- Métricas de Precisión y Error:

REF:	**a las tres**	**y**	**siete**	**minutos de**	**mañana**
HYP:	**a las tres**		**diecisiete**	**minutos de**	**la mañana**

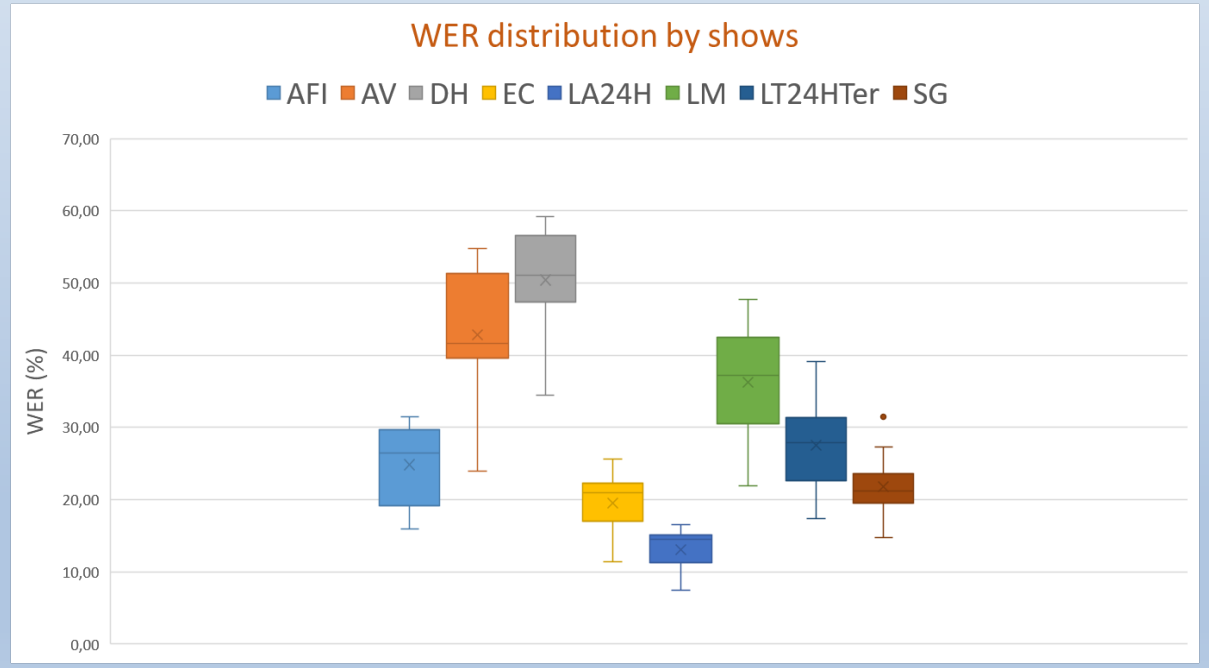
CORRECTO (C)
ERRORES:
Substitutiones (S), Borrados (B), Inserciones (I)

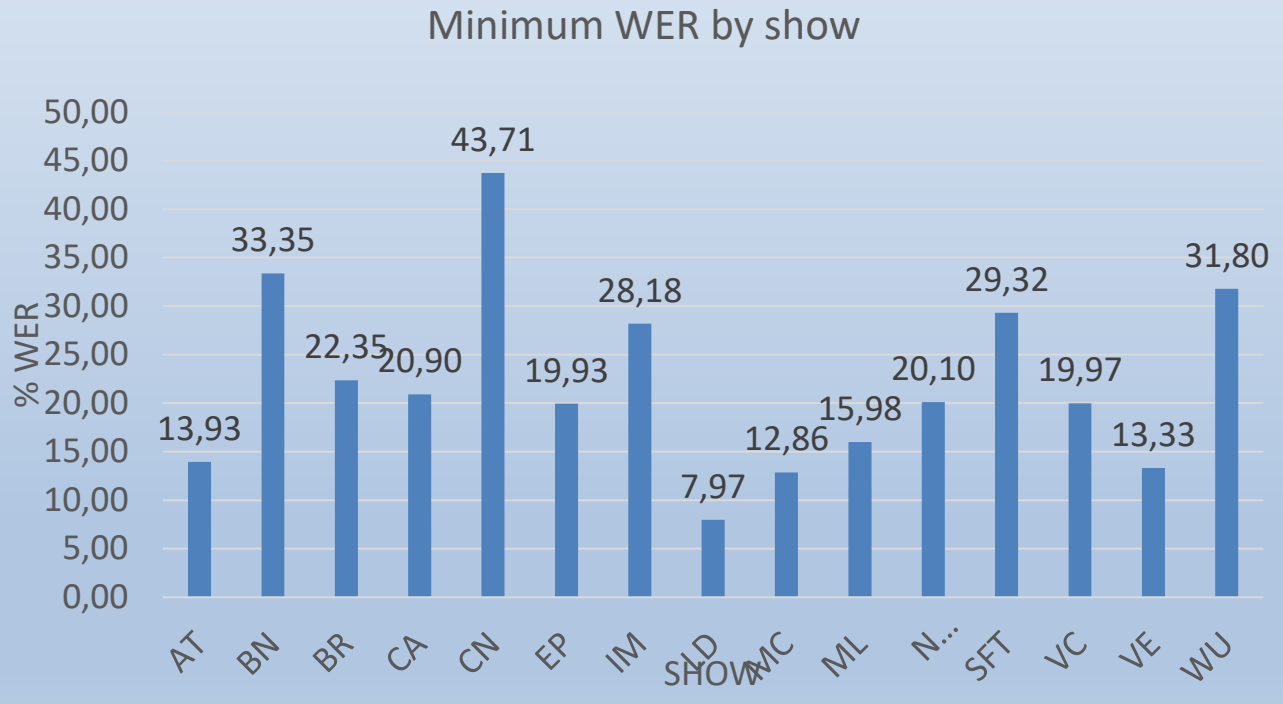
$$\%\ ACC = \frac{C}{C+S+B+I} \times 100$$

$$\%\ WER = \frac{S+B+I}{C+S+B} \times 100$$

Prestaciones: Albayzin 2018 y 2020

- *Iberspeech-RTVE-Challenge* :





Ejemplos de audios de evaluación:

ESPAÑA EN COMUNIDAD

del río hay muchísima la margarita africana de las entre las dunas y bueno así muy bonita y la uña de gato la red natura dos mil a la que pertenece a la playa de fresh urfé es una red europea de espacios protegidos diseñada para asegurar la supervivencia a largo plazo de las especies y los hábitats naturales en el territorio europeo además esta playa toda la zona costera de navia y el río están incluidas en la zona especial de conservación y de protección para las aves

LA TARDE EN 24H TERTULIA

al otro lado está esperanza aguirre fuera de la política también todos sabemos porque ignacio gonzález no está no estaba en la foto en fin para cuando la responsabilidad políticas hasta donde tiene que llegar el nivel del lodazal de lo que estamos viendo para que los responsables políticos últimos den la cara y asuman la responsabilidad que les toque es verdad que estamos todavía en una fase en este caso concreto de la gürtel valenciana sólo de juicio

Diarización e Identificación de hablantes, Añadiendo Reconocimiento del Habla:

J. Biden: Contenido de la intervención 1 … <Tcomienzo1> <Tfin1>

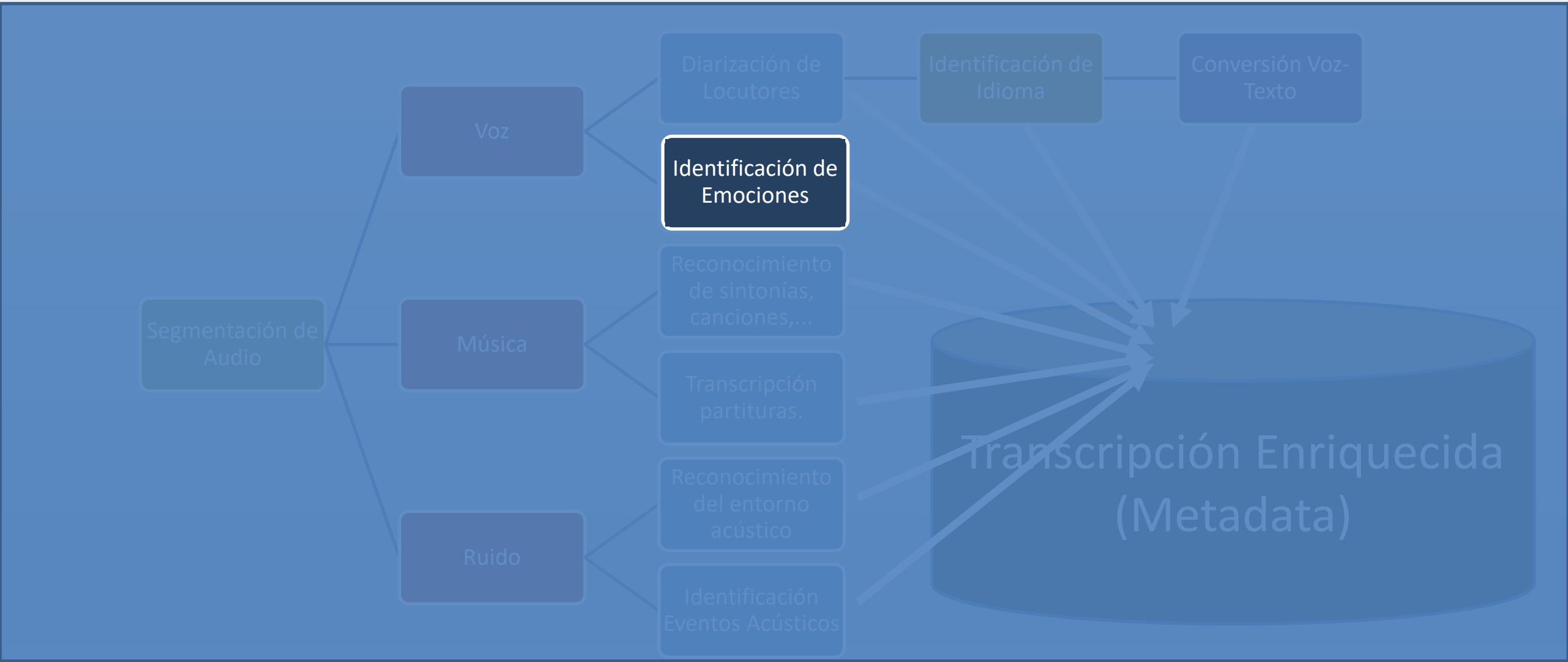
O. Scholz : Contenido de la intervención 2 … <Tcomienzo2> <Tfin2>

L. Truss: Contenido de la intervención 3 … <Tcomienzo3> <Tfin3>

J. Biden: Contenido de la intervención 4 …. <Tcomienzo4> <Tfin4>

A. Scholz : Contenido de la intervención 5 … <Tcomienzo5> <Tfin5>

Tecnologías

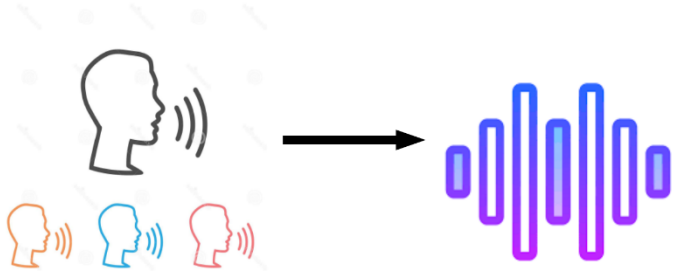


Identificación de Emociones

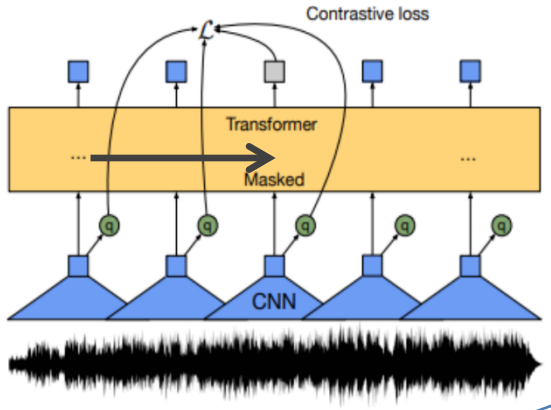
- **¿Para qué sirve?:**
  - **Puede añadir información extra que enriquece el discurso de los protagonistas de un contenido**

Identificación de Emociones

SEÑAL DE VOZ



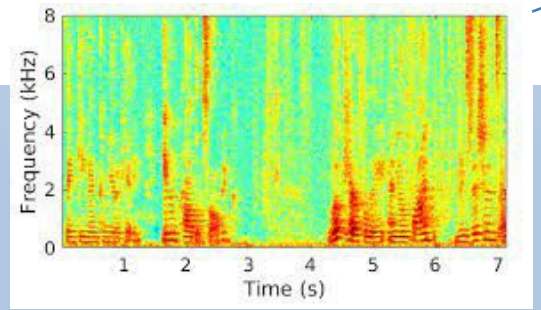
Feature Extraction Using Wav2Vec2

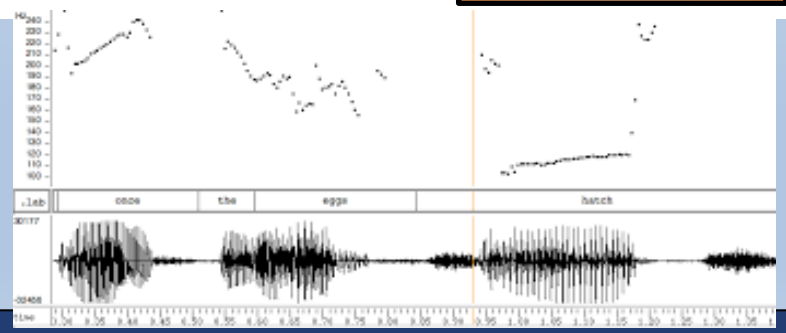


Características:

Espectrales
Prosódicas
Paralingüísticas

...





Classify Layer

Feature Vectors

CLASIFICACIÓN

Emotion Recognition

- **enfado**
- **alegría**
- **tristeza**
- **neutro**
- **aburrimiento**
- **ansiedad**

VIVOLAB

Universidad Zaragoza
1542

Universidad
Zaragoza